

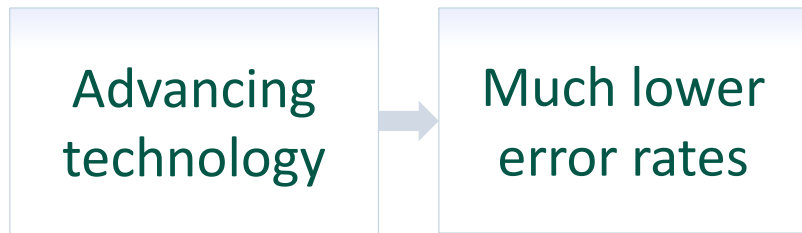


The face recognition company

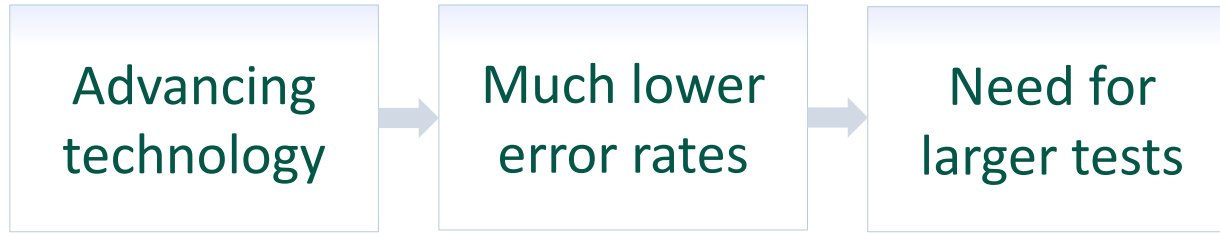
## Effects of Wrong ID Labels

Thorsten Thies

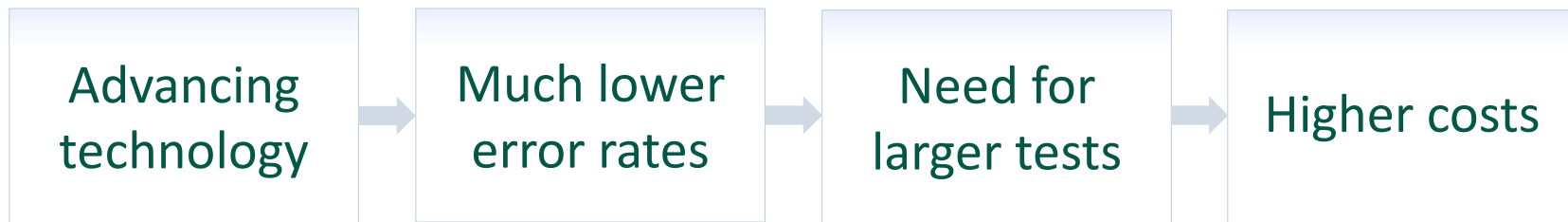
# The story behind this talk



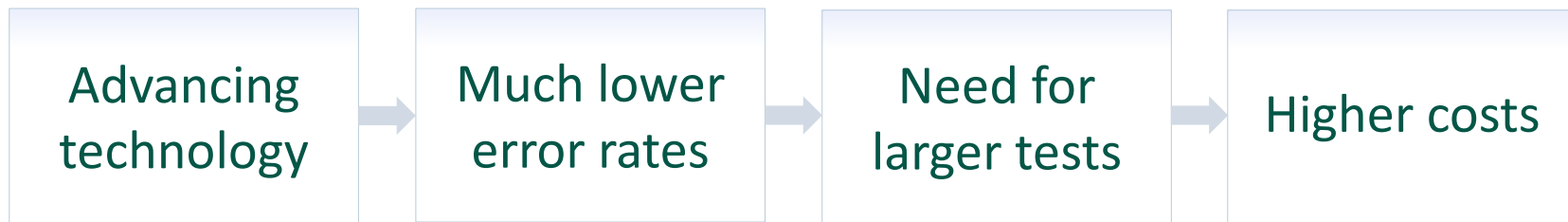
# The story behind this talk



# The story behind this talk



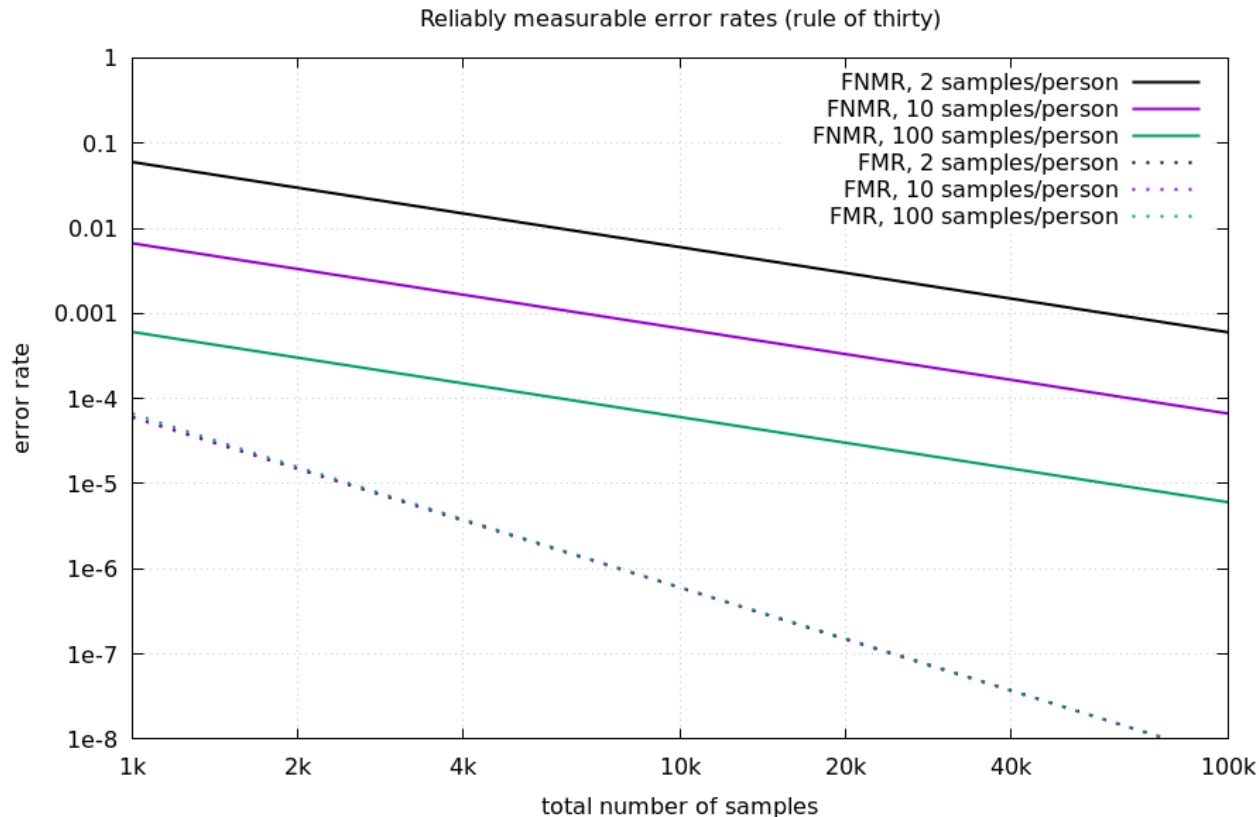
# The story behind this talk



How to measure low FMRs and FNMRs with a limited number of samples?

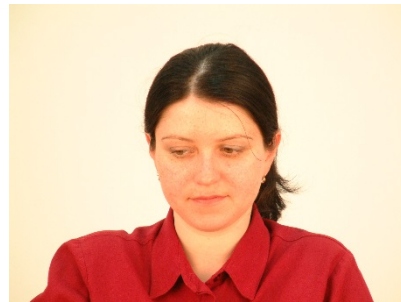
**Solution:** Use a data set with **many samples per person** and do a cross-comparison.

# Which error rates can be measured?



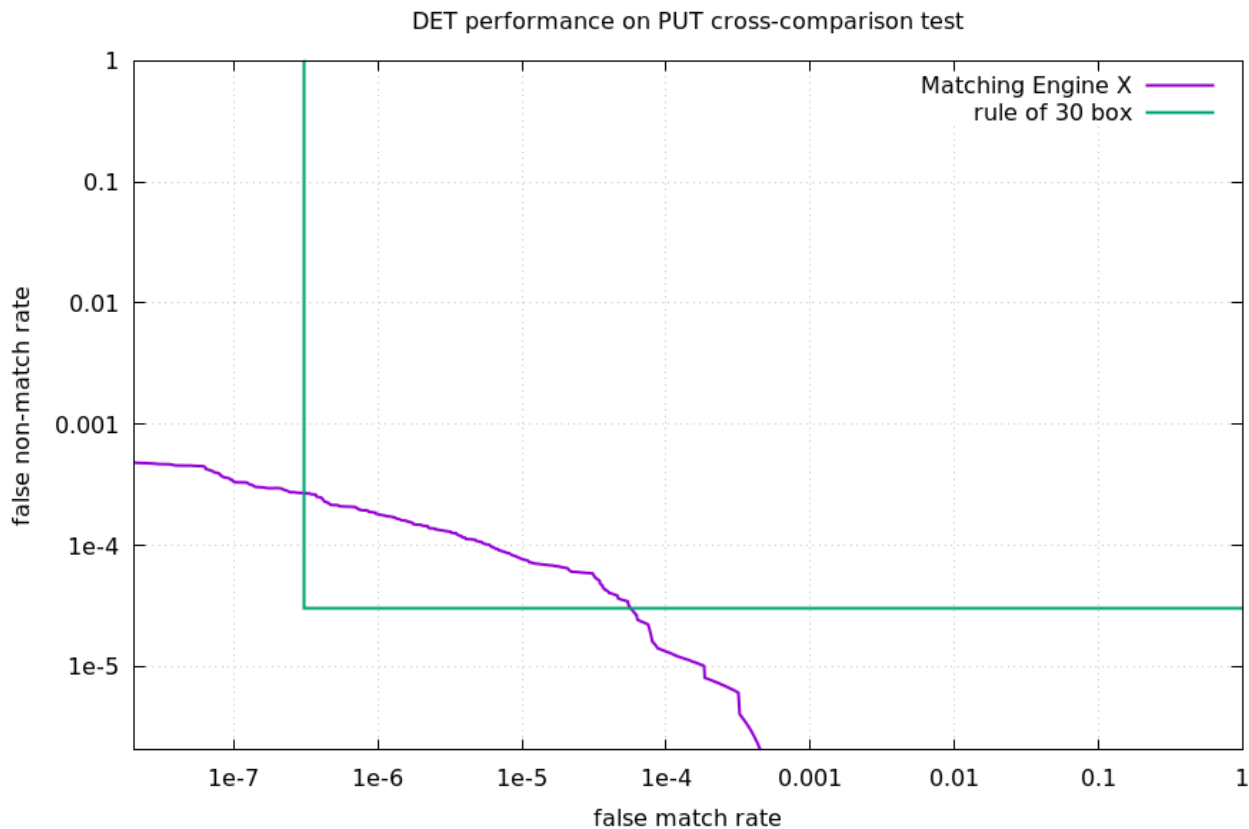
Rule of 30:  
ISO/IEC FDIS 19795-1  
*Information technology -  
Biometric performance  
testing and reporting,*  
annex B.1.2

# PUT face database



9971 face  
images of  
100 persons  
(~100 images  
per person)

# Test result – DET curve



ID1



ID2



ID3



ID4



ID5



ID6



...

ID100



ID1

ID2

ID3

ID4

ID5

ID6

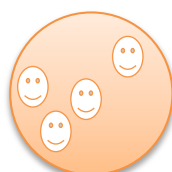
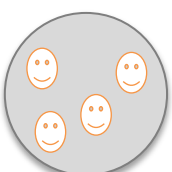
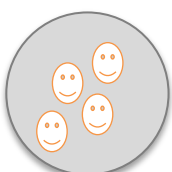
ID100



...



randomly select 50 of the 100 persons



ID1

ID2

ID3

ID4

ID5

ID6

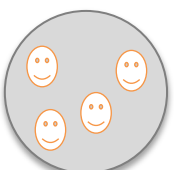
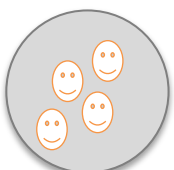
ID100



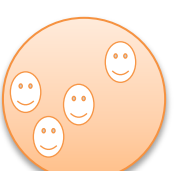
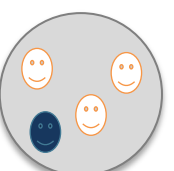
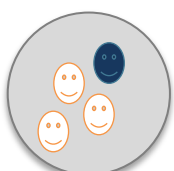
...



randomly select 50 of the 100 persons



for each selected person, randomly pick one sample



ID1

ID2

ID3

ID4

ID5

ID6

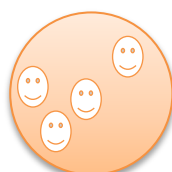
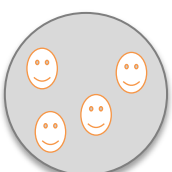
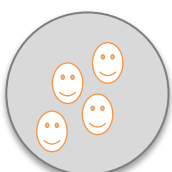
ID100



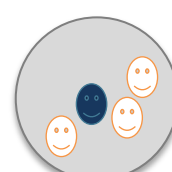
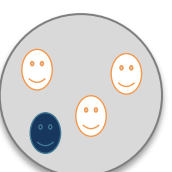
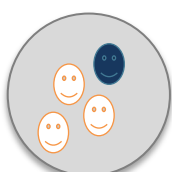
...



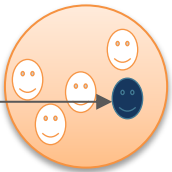
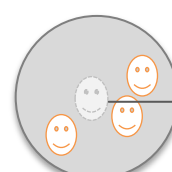
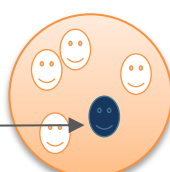
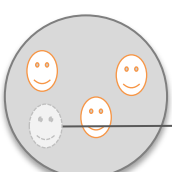
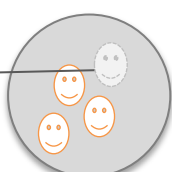
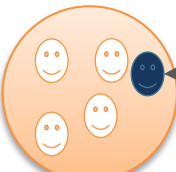
randomly select 50 of the 100 persons



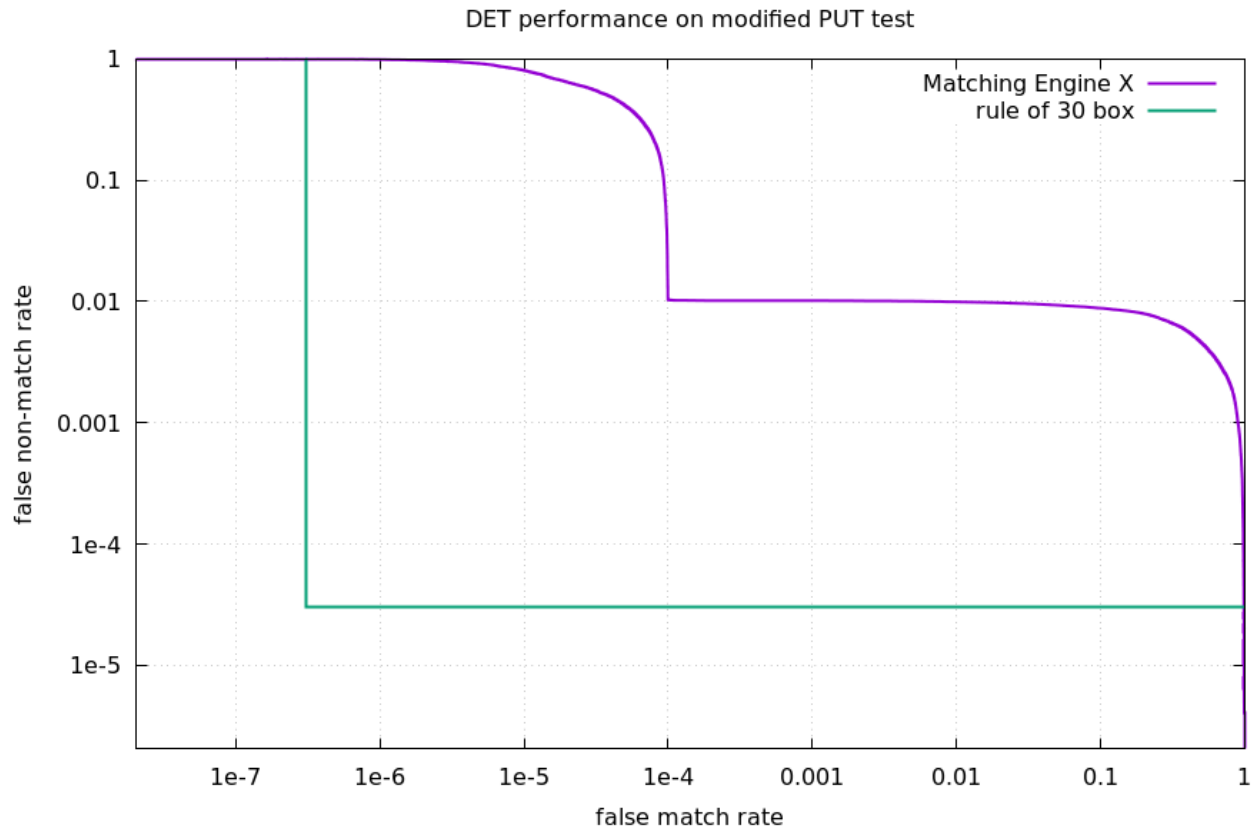
for each selected person, randomly pick one sample



assign the selected samples to another person



# Test result – DET curve



# Slight modification – Strong impact

We altered only 0.5% of the samples.

# Slight modification – Strong impact

We altered only 0.5% of the samples.

*But*

Altering one label affects ~100 positive and ~100 negative scores.

# Slight modification – Strong impact

We altered only 0.5% of the samples.

*But*

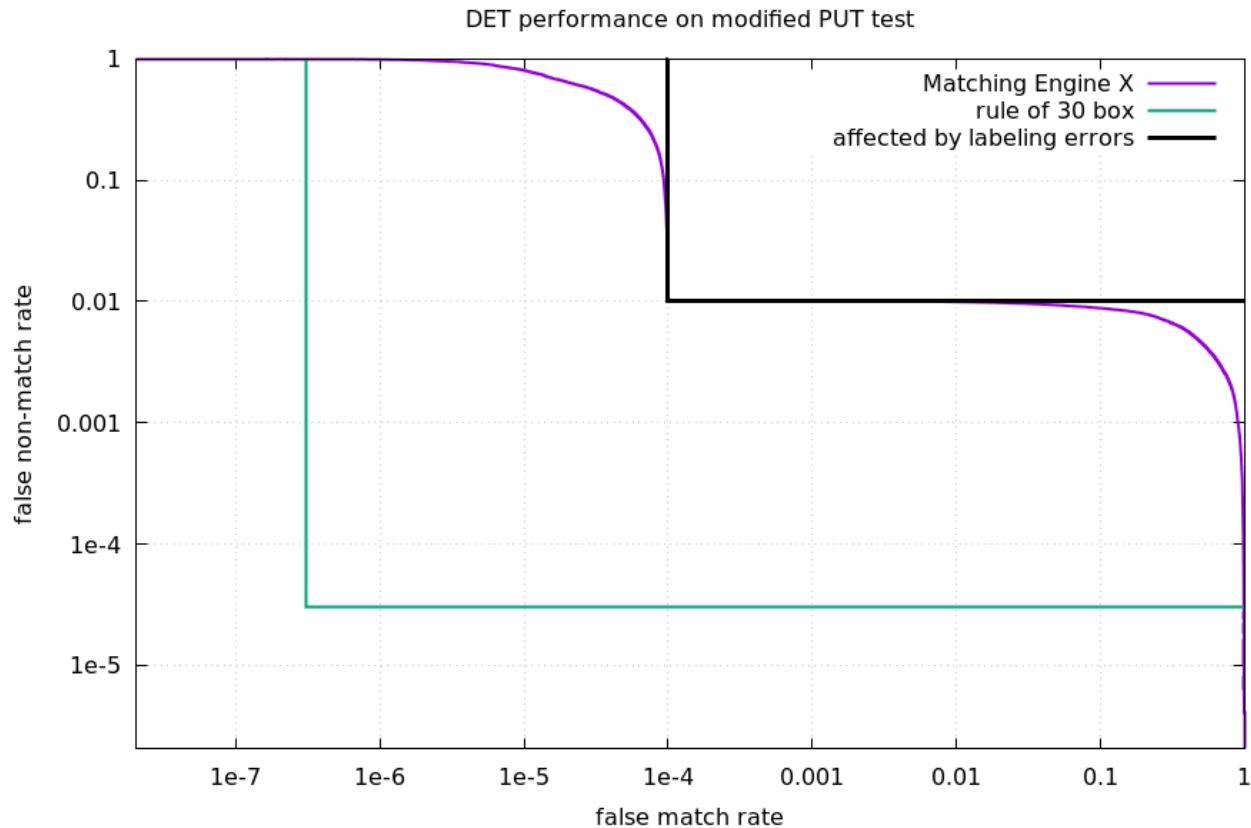
Altering one label affects ~100 positive and ~100 negative scores.



1% of the positive scores are affected.

0.01% of the negative scores are affected.

# Test result – DET curve



# Wrong labels do occur in practice

- different spellings of names / ID labels
- file naming errors
- errors during capturing process
- errors during ID labeling process
- fraud
- . . .

**Note:** Large facial image databases collected from the internet often contain many wrong ID labels.

# Detecting a wrongly labeled sample $x$

**Naive approach:** check positive pairs with scores  $<$  threshold  $T$

# Detecting a wrongly labeled sample $x$

**Naive approach:** check positive pairs with scores  $<$  threshold  $T$

**Drawbacks:**

If  $x$  has all positive scores  $\geq T$ , you won't detect  $x$

If  $x$  has a positive score  $< T$ , there are often *many other positive pairs involving  $x$* , at score  $< T$ , spamming your list

# Detecting a wrongly labeled sample $x$

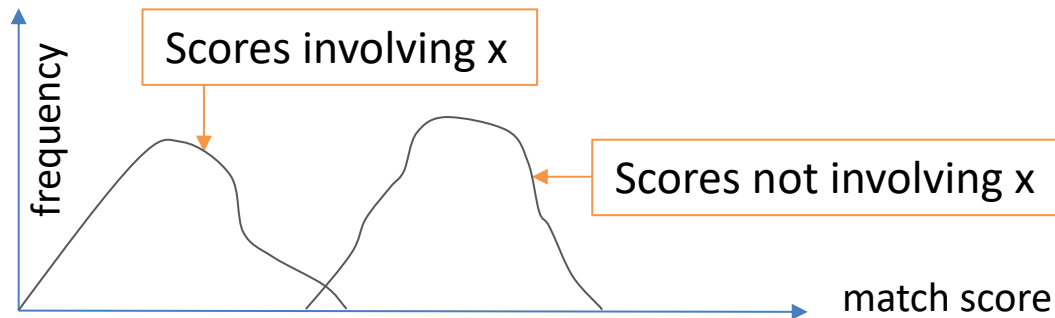
**Naive approach:** check positive pairs with scores  $<$  threshold  $T$

**Drawbacks:**

If  $x$  has all positive scores  $\geq T$ , you won't detect  $x$

If  $x$  has a positive score  $< T$ , there are often *many other positive pairs involving  $x$* , at score  $< T$ , spamming your list

**Better:** consider *all* positive scores involving  $x$ , at once



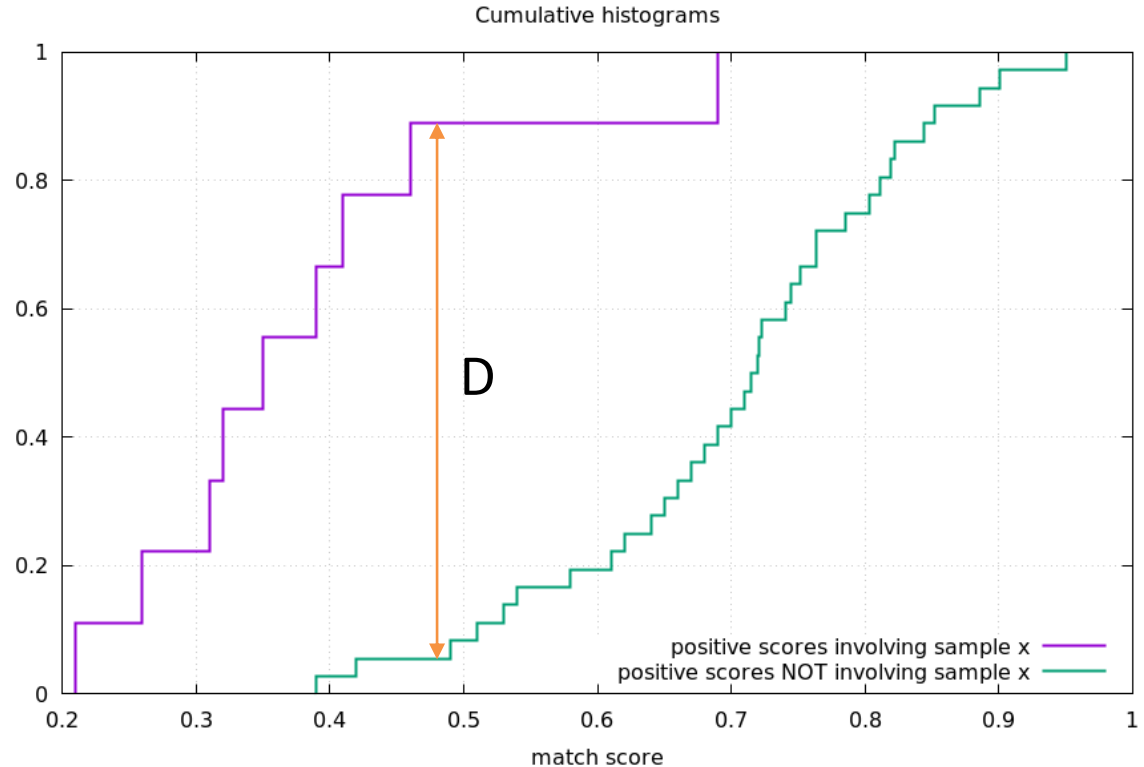
# Kolmogorov-Smirnov Test

compares two  
distributions,  
of N resp. M scores

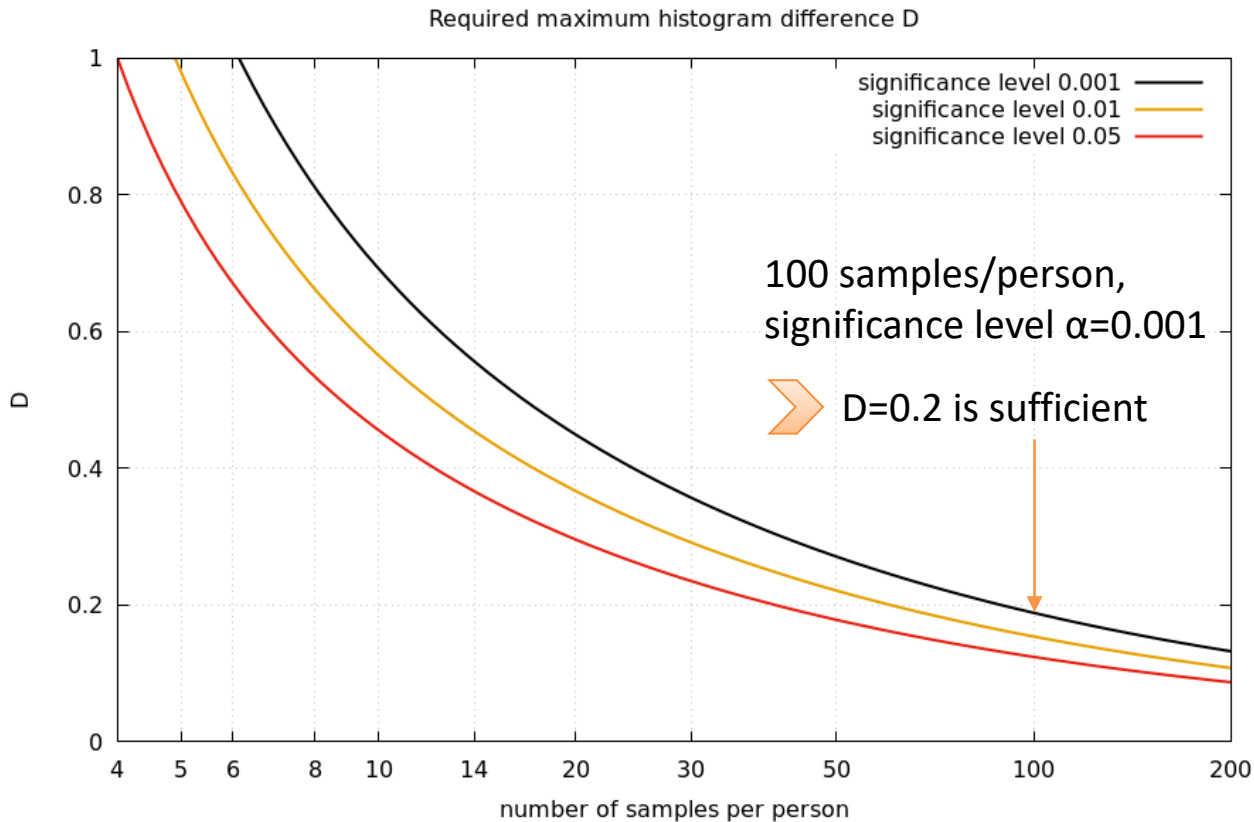
rejects the  $H_0$ -hypothesis  
(„scores stem from the  
same distribution“)  
at level  $\alpha$  if

$$D > \sqrt{-\frac{1}{2} (\log \alpha) \frac{N+M}{NM}}$$

M. Hollander, D. Wolfe, E. Chicken,  
*Nonparametric statistical methods*,  
3. ed., Wiley (2013)



# How large does D need to be?



# Outlier detection method

	x1	x2	x3	x4	...	xn
x1	1	.97	.42	.89	...	.98
x2	.97	1	.31	.79	...	.99
x3	.42	.31	1	.62	...	.15
x4	.89	.79	.62	1	...	.82
...	...	...	...	...	...	...
xn	.98	.99	.15	.82	...	1

Score matrix of all  $n$  samples of a person

# Outlier detection method

x →

	x1	x2	x3	x4	...	xn
x1	1	.97	.42	.89	...	.98
x2	.97	1	.31	.79	...	.99
x3	.42	.31	1	.62	...	.15
x4	.89	.79	.62	1	...	.82
...	...	...	...	...	...	...
xn	.98	.99	.15	.82	...	1

Pick a sample x and mark all related scores

# Outlier detection method

x →

	x1	x2	x3	x4	...	xn
x1		.97	.42	.89	...	.98
x2	.97		.31	.79	...	.99
x3	.42	.31		.62	...	.15
x4	.89	.79	.62		...	.82
...	...	...	...	...		...
xn	.98	.99	.15	.82	...	

Ignore irrelevant  
identical comparisons

# Outlier detection method



x →

	x1	x2	x3	x4	...	xn
x1						
x2	.97					
x3	.42	.31				
	.89	.79	.62			
...	...	...	...	...		
xn	.98	.99	.15	.82	...	

Ignore redundant  
symmetrical scores



# Outlier detection method

	x1	x2	x3	x4	...	xn
x1						
x2	.97					
x3	.42	.31				
	.89	.79	.62			
...	...	...	...	...		
xn	.98	.99	.15	.82	...	

- determine the cumulative histograms for each sample  $x$ :
- $A(x)$  of positive scores involving  $x$  
- $B(x)$  of positive scores NOT involving  $x$  

# Outlier detection method

	x1	x2	x3	x4	...	xn
x1						
x2	.97					
x3	.42	.31				
	.89	.79	.62			
...	...	...	...	...		
xn	.98	.99	.15	.82	...	

- determine the cumulative histograms for each sample  $x$ :
- $A(x)$  of positive scores involving  $x$  
- $B(x)$  of positive scores NOT involving  $x$  
- compute maximum absolute difference  $D$  between  $A(x)$  and  $B(x)$

# Outlier detection method

	x1	x2	x3	x4	...	xn
x1						
x2	.97					
x3	.42	.31				
	.89	.79	.62			
...	...	...	...	...		
xn	.98	.99	.15	.82	...	

- determine the cumulative histograms for each sample x:

- A(x) of positive scores involving x 

- B(x) of positive scores NOT involving x 

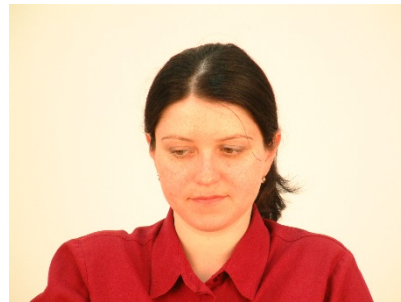
- compute maximum absolute difference D between A(x) and B(x)

- compute p-value:

$$p(D) = \exp \left( -2 D^2 \frac{(n-1)(n-2)}{n} \right)$$

- for a threshold T, report all x with  $p(D) < T$ , ranked by p(D)































































# PUT face database



9971 face  
images of  
100 persons  
(~100 images  
per person)

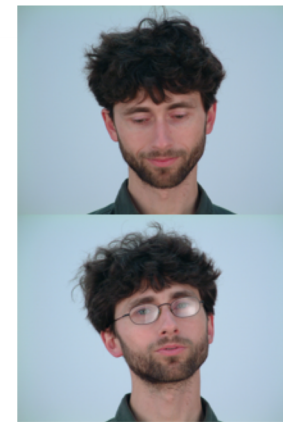
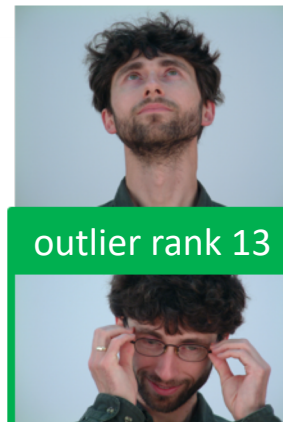
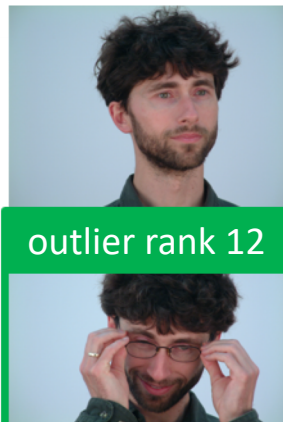
IDs of 50 samples changed

# Results

Rank 1-10										
11-20										
21-30										
31-40										
41-50										
51-60										
61-70										
71- ...									...	

**All 50** wrongly  
labeled samples  
appear among the  
**top 53** outliers.

# Three outliers with *correct* ID label



# Know your algorithm – and your data

Running this outlier detection makes sense even if your data has entirely correct labels:

- It can point you to unusual samples in your data, e.g. image capturing failures.
- It indicates which variations among the samples of a person are easier or harder for your algorithm.

# In summary

- 1 Cross-comparison tests with many samples per person are **efficient to reliably measure low error rates**.
- 2 However, they are **sensitive to wrong labels**.
- 3 The **Kolmogorov-Smirnov test** finds wrong label samples and other outliers efficiently.



# The face recognition company

Thank you! Questions?

[www.cognitec.com](http://www.cognitec.com)  
[info@cognitec.com](mailto:info@cognitec.com)

We are committed to delivering the best face recognition performance available on the market.